



Cross-linguistic phonetic recalibration in bilingual lexical processing

Yuhyeon Seo*, Department of Psychological Sciences, State University of New York at Oswego, Oswego, NY, USA, yuhyeon.seo@oswego.edu

Olga Dmitrieva, School of Languages and Cultures, Purdue University, West Lafayette, IN, USA

*Corresponding author.

The study explores phonetic recalibration as a novel method for investigating sound-category linkages across languages, by examining cross-linguistic generalization of phoneme retuning among Spanish–English bilinguals ($n = 110$). Participants were trained to recalibrate the perceptual boundary between /p/ and /k/ in one of their languages, English or Spanish, via a lexical decision task which included words with acoustically manipulated /p/ or /k/. Recalibration was subsequently tested via an identification task in both of their languages, thus determining whether recalibration generalized to the untrained language. The results revealed asymmetric generalization patterns. While recalibration always occurred in the trained language, it transferred to the untrained language only in the direction of English to Spanish, but not Spanish to English, suggesting that L1 and L2 sound categories were asymmetrically linked. We discuss possible explanations for this asymmetry, including the language dominance profile of our participants. These findings expand significantly on previous cross-linguistic recalibration research through two methodological innovations. First, recalibration was tested across phonetically distinct categories: voiceless aspirated stops in English and voiceless unaspirated stops in Spanish. Second, cross-language training and testing stimuli were produced by different speakers, ruling out the possibility that generalization across languages could be attributed to speaker-specific learning.



1. Introduction

The cross-linguistic organization of sound categories in bilinguals, including second language (L2) learners, is a critical factor in bilingual speech production and perception according to the most prominent theoretical approaches to L2 speech, such as the revised Speech Learning Model (SLM-r; Flege & Bohn, 2021), the Perceptual Assimilation Model of L2 speech learning (PAM-L2; Best & Tyler, 2007; Tyler & Best, 2024), and the Second Language Linguistic Perception model (L2LP; Escudero, 2005, 2009; Escudero & Yazawa, 2024). One of these theories, SLM-r, specifically posits that the formation of L2 sound categories that are distinct and independent from the first language (L1) sound categories constitutes an important milestone in the acquisition of L2 speech, and a predictor of target-like production and perception of L2 sounds.

Theoretical approaches differ in terms of the levels of representations with which they operate. Both SLM (Flege, 1995, 2003) and SLM-r postulate that L1-L2 interaction occurs at the phonetic level of position-specific allophones, while PAM-L2 and L2LP involve both the phonetic and the phonological levels. Nevertheless, all three approaches congruently posit that L2 speech learning essentially consists of creating distinct L2 sound categories, be they phonetic or phonological, perceptual, acoustic, or articulatory. Furthermore, it is generally agreed that the phonological and/or phonetic correspondences (or lack thereof) between the two sound systems determine the likelihood and speed of L2 category formation.

SLM-r addresses both perception and production in L2, using the patterns of perceptual assimilation among L1 and L2 categories to make predictions regarding L2 production, unlike PAM-L2 and L2LP, which focus on L2 perception. The SLM-r approach explicitly incorporates the perception-production link—a fundamental condition of speech communication. In cases of cross-linguistic perceptual similarity, SLM-r predicts initial difficulty in creating L2 categories.

Specifically, prior to creation of independent L2 categories, L2 sounds are linked to the perceptually similar L1 categories and together form composite L1-L2 interlingual categories. Both PAM-L2 and L2LP make conceptually similar assumptions regarding the difficulty of perceptually learning new L2 categories, which in turn predict their perceptual discrimination from other L2 sounds. SLM-r takes it one step further and predicts that, as a result of such ‘equivalence classification,’ a mutual attraction between the linked L1 and L2 phonetic categories is expected in production, such that they acoustically assimilate to each other. In contrast, once an independent L2 category is established, the two may dissimilate from each other, becoming more acoustically distinct. Therefore, synthesizing, in general terms, the assumptions of the three models, L2 learners’ ability to perceptually learn sound categories in the L2 predicts both perception and production of these L2 sounds, as well as perception and production of related L1 categories.¹

¹ Although all three models acknowledge the possibility of L2 effects on L1 production or perception, they attribute these effects to different mechanisms. SLM-r and PAM-L2 derive these effects from direct interaction of sound categories in the shared interlanguage system, while L2LP attributes them to the parallel activation of the independent phonological grammars of the two languages.

Existing approaches to investigating the structure of L1/L2 phonetic space primarily have relied on production data as evidence for L2 category formation. Comparing bilinguals' productions of L1 and L2 phones to those of monolingual speakers of both languages is often employed to substantiate category formation (e.g., Baker & Trofimovich, 2005; Casillas, 2020; Sancier & Fowler, 1997, among others). For example, Flege et al. (2003) argued that Italian immigrants in Canada who produced English [eɪ] as significantly more monophthongal than native English speakers did not form a separate L2 category for this diphthong category, as evidenced by its assimilation to Italian [e]. In the same study, the authors argued that another group, which demonstrated a dissimilatory relationship between Italian [e] and English [eɪ] by producing English [eɪ] with exaggerated formant movements, compared to native English speakers, formed a separate [eɪ] category.²

A consistent issue with interpreting such cross-linguistic phonetic interactions (assimilation or dissimilation) as evidence of L2 category formation (or lack thereof) is that no independent and generally accepted method to investigate the structure of cross-linguistic phonetic space exists (see, for example, Flege & Bohn, 2021, p. 41: "A method did not exist in 1995 for determining when a new L2 phonetic category had been formed and, alas, the same holds true today"). In the absence of such a method, the reasoning becomes circular: Cross-linguistic phonetic interactions are explained by category formation and are used as evidence for category formation.

In this study, we explore a methodology that can be adapted for determining whether sound categories are separate or integrated across languages in bilingual speakers, specifically the methodology of *lexically guided phonetic recalibration* or *phoneme retuning*. The use of this conceptually different methodology allows us to circumvent the issue of circular reasoning: the relationship between the sound categories of the L1 and L2 is determined without relying on cross-linguistic phonetic interactions. Therefore, the theoretical stipulations ascribing specific L2 production or perception behaviors (assimilation to or dissimilation from related L1 sounds) to L2 sound category formation can be verified by establishing, using this method, that the L2 category is in fact independent from the L1. Beyond corroborating theoretical assumptions of SLM-r and competing approaches, such a methodology could provide a general insight into the structural organization of sound categories in bilinguals. Since sound categories are cognitive entities that cannot be observed directly, we aim to determine whether sound categories are linked across languages by exerting an influence on the perceptual boundaries of a phonemic category in one language and observing whether a similar change is registered for its counterpart in another language. The presence of such a link is interpreted as evidence that the L1 and

² In the perceptual domain, L2 category formation has been supported by evidence of shifts in discrimination peaks between L2 categories (Guion et al., 2000; Thorin et al., 2018), as well as by findings of native-like categorization of L2 sounds—though native-likeness is often defined in an ad hoc manner (e.g., Takahashi, 2023) or inferred from category mapping patterns (e.g., Cutler et al., 2006).

corresponding L2 categories are not fully independent of each other.³ We do not make strong a priori assumptions regarding the phonetic or phonological nature of sound categories that interact across languages. However, such assumptions make different predictions regarding the outcomes of the experiment, which we discuss in section 3.5.

2. Literature review

2.1. Phoneme retuning across languages

Lexically guided retuning of phonemic categories is an example of perceptual learning that is critical to our ability to adapt to variability in speech, including novel dialects, accents, and individual pronunciation differences. When listeners are exposed to sounds with nonprototypical acoustic profiles, for example, fricatives that are acoustically intermediate between /f/ and /s/, embedded into disambiguating lexical items, e.g., *ungrateful*, they learn to categorize such instances as exemplars of /f/, thus expanding the boundaries of the phonemic /f/ category. This reorganization can be detected via a forced-choice categorical decision test whereby more items on the acoustic continuum between /f/ and /s/ are categorized as /f/, as a result of lexically guided perceptual learning. Such findings are well established in single-language experiments with L1 (presumably largely monolingual) speakers of the language (Kraljic & Samuel, 2006; Norris et al., 2003; Reinisch et al., 2014). Only recently, researchers began examining lexically guided recalibration in participants' non-L1 languages and across languages in bilinguals. A very limited body of work presently indicates that (a) recalibration is not limited to listeners' L1 and (b) recalibration can generalize across languages, such that learning to expand phonemic categories in one language (L1 or L2) results in a similar expansion for the comparable category in bilinguals' other language (L2 or L1). For example, Reinisch et al. (2013) showed that recalibration of /s/ and /f/ categories occurred for L1 speakers of German trained and tested on Dutch, their L2. The same study demonstrated that Dutch listeners recalibrated /s/ and /f/ categories in the Dutch test after training on the Dutch-accented English of the same speaker who provided the test recordings. Importantly, the authors interpreted these findings as evidence that listeners considered non-prototypical pronunciation of /f/ and /s/ to be the intrinsic characteristic of the speaker, thus generalizing this attribute across the languages spoken by the same person. The authors also noted that given the highly similar acoustic properties of /f/ and /s/ across Dutch and German languages, the establishment of language-specific categories is likely unnecessary.

Schuhmann (2016) examined cross-linguistic phonetic recalibration of /s/ and /f/ in L1 German learners of English and L1 English learners of German. Both groups were trained on English but tested on both English and German. Phonetic recalibration was observed in all four cases, including L1 English speakers tested on L2 German and L2 English speakers tested on L1 German,

³ While we do not assume that a complete separation of L1 and L2 sound categories is either necessary or ultimately achievable in L2 learning, such a stage is consistent with the developmental trajectory proposed by the SLM-r framework.

where the training language and the test language did not match. Given that the effect sizes of perceptual recalibration were not identical across exposure and test language conditions, the author concluded that L1 and L2 categories were linked but not merged (separate but interconnected) across languages. Of note again is the fact that /s/ and /f/ are acoustically highly similar in English and German, thus their strong cross-linguistic link is to be expected. Another important aspect of the design is that the same speaker recorded both the training and the testing stimuli in English and German, which may have induced speaker-specific learning effects. Finally, the test items were nonwords, where only the quality of /r/ (approximant or trill) indicated the ‘language’ of the stimuli, making it difficult to evaluate the generalizability of the findings to real words.

Recently, Caudrelier et al. (2023) and Caudrelier et al. (2024) investigated the recalibration of /s/ and /f/ in French-English bilinguals in Montreal, Canada. As observed in the previous two studies, recalibration was found to transfer from the training language (English or French) to the unmatched test language (French or English), but the recalibration effect was smaller in the cross-language than in the within-language condition. The authors concluded that bilingual listeners may have shared /f/ and /s/ phonemic categories across their two languages, but the links may be stronger across closely related (Germanic) languages than in the case of Romance-Germanic language combination. The study also adopted a single-speaker design: The training and test recordings were provided by the same French-English bilingual speaker, raising the possibility of listeners accommodating to a specific speaker, rather than generalizing abstract perceptual learning across languages.

What all these studies share is the exclusive use of /s/ and /f/ for cross-linguistic recalibration, while recognizing that these phonemes are highly phonetically similar across the languages under investigation. Additionally, all previous studies adopted a single-speaker design in recording audio stimuli, with the same speaker recording both the exposure and the training stimuli. Because of these methodological choices, their findings, while possibly indicating shared phonemic categories across languages, could also be due, at least in part, to speaker-specific perceptual learning effects. Yet lexically guided recalibration is not limited to speaker-specific scenarios. For instance, Kraljic and Samuel (2006) showed that recalibration trained on /d/ and /t/ generalized to another speaker and a parallel phonemic contrast, /b/ and /p/, indicating that the process can be speaker-independent and feature- rather than phoneme-based (see also Llompart, 2024, for cross-talker generalization in perceptual retuning of vowels by advanced learners of English). Importantly, Kraljic and Samuel (2006) suggested that the use of fricatives, which are known to provide rich speaker-specific information, can incentivize phoneme-specific and speaker-specific learning (although Kraljic and Samuel (2005) also show partial speaker generalization with fricatives). To minimize the speaker-specific learning effects, they suggested the use of stops, given the weaker speaker-specific spectral cues, compared to fricatives (Repp, 1984). This hypothesis was further supported in Kraljic and Samuel (2007), who reported speaker-specific recalibration for fricatives but more generalized recalibration for stops. In the present study, we recruited different speakers to record the exposure (training) and the testing stimuli

across languages. Both our speakers were female, ensuring acoustic similarity that is believed to facilitate cross-speaker generalization (Kraljic & Samuel, 2005; Tamminga et al., 2020). We also investigated recalibration of the place contrast in stop consonants instead of fricatives to mitigate speaker-specific learning effects.

The choice to use stops allows us to address an additional question. The consensus in the literature on bilingual phonetics and phonology is that for L2 categories that are highly phonetically similar to their L1 counterparts, to the point of being perceptually indistinguishable, creation of distinct categories is either impossible or unnecessary (e.g., PAM-L2; Best & Tyler, 2007). Thus, the inquiry into cross-linguistic phonetic influence has focused on pairs of sounds that are phonetically similar but non-identical across languages, such as aspirated voiceless stops in English versus unaspirated voiceless stops in Spanish. For such pairs, their cross-linguistic relationship was theorized to be more dynamic and subject to developmental changes, including eventual, but not immediate, creation of distinct L2 categories (SLM-r; Flege & Bohn, 2021). For cross-language pairs such as these, the generalization of lexically guided recalibration from the training language to a different test language is not necessarily a foregone conclusion. The use of this methodology allows us to investigate whether phonetically similar but non-identical categories are linked or integrated across languages to such an extent that transfer of phonetic recalibration from one language to another can occur. It is important to note that given the hypothesized dynamic nature of the cross-linguistic relationship between such sounds, the findings may differ depending on the circumstances of bilingualism and the stage of L2 acquisition.

2.2. The present study

The present study investigates lexically guided recalibration of perceptual boundaries between the voiceless stops /p/ and /k/ in Spanish-English bilinguals residing in the United States. Exposure to sounds acoustically intermediate between /p/ and /k/ was provided via a lexical decision task, and resulting recalibration was tested via a two-alternative forced choice (2AFC) categorization task with real words of English and Spanish. Two different L1 speakers recorded English and Spanish stimuli. Bilingual participants were recruited online and randomly assigned to one of the four groups: Two groups were trained in English, and two in Spanish. Within each language group, some participants were exposed to ambiguous /p/, while others were exposed to ambiguous /k/. Regardless of training language, all participants completed the 2AFC identification task in both English and Spanish. As a result, for two of the groups, the training and test language matched (with the same speaker providing the training and the test recordings), whereas for the other two groups, the training and test languages differed (with a different speaker providing the training and the test stimuli). The research question we are addressing with this experimental design is the following: Do bilinguals who were trained to shift the perceptual boundaries of the place contrast between /p/ and /k/ in one of their languages (e.g., English) demonstrate resulting recalibration in both languages, Spanish and English?

Crucially, voiceless stops in English are aspirated in word-initial position and at the onset of syllables with any degree of stress, while in Spanish voiceless stops are unaspirated across the board (Lisker & Abramson, 1964). Thus, for the transfer of recalibration to occur, bilinguals must categorize Spanish voiceless *unaspirated* and English voiceless *aspirated* stops together, despite their phonetic differences. We hypothesized that the transfer of recalibration would be observed across languages. This hypothesis was motivated by frequent findings in previous literature that voiceless stops across languages, such as English and Spanish, are subject to cross-linguistic influence in the assimilatory direction, although assimilation can be asymmetric as the more dominant or earlier acquired language tends to exert a stronger influence on the less dominant or later acquired language (e.g., Flege & Eefting, 1987, among others). These findings suggest that voiceless stops can often be linked or integrated across languages in bilingual cognitive grammar.

An important question, however, is what exactly is linked cross-linguistically: context-specific allophones of voiceless stops or abstract phonological representations of voiceless stops. SLM-r assumes the former, while PAM-L2, the latter. Finally, L2LP upholds that L1 and L2 perceptual grammars are separate and independent from each other. These assumptions make different predictions, given the methodology of our investigation, as discussed in the following section.

3. Methods

3.1. Participants

A total of 127 Spanish (L1)-English (L2) bilingual speakers participated in the study. They were recruited and compensated on Prolific and performed experimental tasks via Gorilla (Anwyl-Irvine et al., 2020).⁴ Among these, 17 participants yielded unreliable data in the categorization task, and their data were excluded from analysis, resulting in a total of 110 participants (Age $M = 29.5$, Age range = 18–48). We identified unreliable data using a combination of statistical analysis (mixed-effects logistic regression on categorization responses for each participant) and visual/manual inspection of response patterns and reaction times. Participants were excluded if their responses were not significantly affected by Step (an acoustic continuum between /p/ and /k/), as this indicated that they were not systematically influenced by the stimuli continuum from /p/ to /k/. Participants were also excluded if they selected the same response at least five consecutive times, despite variation in Step, and if this pattern occurred more than once.

All retained participants reported acquiring Spanish first and English later. All participants were residing in the US at the time of the experiment. The majority of participants could be categorized as heritage speakers born and raised in the US ($n = 83$), while others were early

⁴ Our experimental design required participants to attend to both the auditory and the orthographic presentation of the stimuli in a relatively quiet environment, but there was no mechanism to ensure that every participant followed the instructions faithfully since participants were recruited online. While laboratory studies are not fully immune to such issues, we acknowledge this vulnerability in online data collection.

bilinguals who immigrated before the age of eight ($n = 9$) or late bilinguals who immigrated after the age of eight ($n = 18$).⁵

To assess the linguistic backgrounds of the participants, the present study employed an adapted version of the Bilingual Language Profile questionnaire (BLP, Birdsong et al., 2012). Specifically, we excluded the language attitude module of the BLP, as it has been argued to be a less reliable indication of language dominance compared to other modules due to cultural differences in language attitudes (see Olson, 2023, for more discussion). Therefore, the questionnaire included three modules, including language history, use, and proficiency, yielding bilingual dominance scores ranging from -164 (English-dominant) to 164 (Spanish-dominant), with the score of zero representing balanced bilingualism. The BLP score is calculated by summing up the scores in each section with equal weights.

Participants' dominance scores, as measured by the modified BLP, are shown in **Figure 1**. The mean score was -14.3 ($SD = 34.9$), suggesting that on average participants were slightly English-dominant. A Bayesian one-sample t -test, with the μ value of 0 indicating balanced bilingualism as the test value, showed that participants were highly unlikely to be balanced bilinguals ($BF_{10} = 421.4$ ⁶). Indeed, individuals varied considerably with the scores ranging from -95.9 (English-dominant) to 80.7 (Spanish-dominant).

Each section of the BLP showed response patterns consistent with the overall dominance scores. In the subsections of BLP, higher score indicates greater history, more frequent use, and higher proficiency in the corresponding language. The Language History section assessed participants' Age of Acquisition (AoA), Length of Residence (LoR), comfortable language, education language, family language, and work language. History scores for Spanish ($M = 34.5$, $SD = 7.4$) and English ($M = 36.3$, $SD = 8.9$) were not notably different ($BF_{10} = 0.3$), suggesting that the timing and contextual distribution of participants' exposure were relatively balanced across the two languages, but with some degree of variability in each language. For instance, participants differed in how comfortable they felt about using Spanish and English, the years of Spanish or English classes they had taken, LoR, and whether or not they used English or Spanish at work. In contrast, participants were consistent in indicating Spanish as their L1 and their family language. The Language Use section evaluated the frequency of Spanish and English use in different contexts. Use scores for Spanish ($M = 16.2$, $SD = 14.9$) were noticeably lower than those for English ($M = 26.9$, $SD = 18.8$) ($BF_{10} = 7.13 \times 10^5$), which was expected as all participants were residing in the US, although both languages showed substantial variability across participants.

⁵ We acknowledge that some of these *early* bilinguals could also be classified as heritage speakers depending on the definition. In the present study, we operationalize heritage speakers as bilingual speakers born and raised in the US (Seo, 2024; Seo & Dmitrieva, 2024).

⁶ BF_{10} is a notation of Bayes Factor that compares the evidence for the alternative hypothesis (H_1) to the null hypothesis (H_0). The greater the value, the stronger evidence in favor of H_1 over H_0 . The BF_{10} of 421.4 indicates that the data are 421.4 times more likely under H_1 compared to H_0 .

The Language Proficiency section measured self-reported proficiency in comprehension, speaking, reading, and writing. Proficiency scores were decisively higher for English ($M = 51.4, SD = 5.0$) than for Spanish ($M = 44.8, SD = 8.2$) ($BF_{10} = 1.79 \times 10^9$), again with a high degree of individual difference. The BLP reports suggest that the participant pool in this study was diverse in terms of linguistic background, while indicating higher proficiency and use in English than in Spanish. **Table 1** provides a summary of participants' linguistic backgrounds.

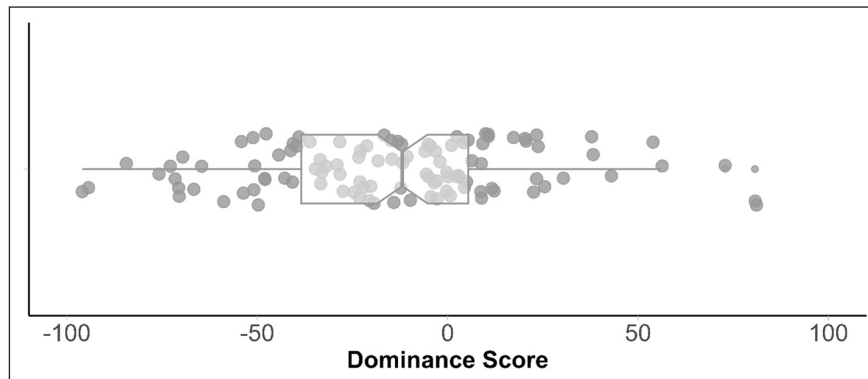


Figure 1: Participants' BLP scores. Positive values indicate Spanish dominance, and negative values indicate English dominance.

Background	$M (SD)$	Range
Age	29.5 (7.0)	18–48
Age of Immigration to the US	3.5 (7.5)	0–31 ⁷
English AoA	5.3 (3.2)	0–18 ⁸
Spanish AoA	0.4 (1.0)	0–5
Dominance (BLP)	–14.3 (34.9)	–95.9–80.7
BLP Subsections	Language	$M (SD)$
Language History	Spanish	34.5 (7.4)
	English	36.3 (8.9)
Language Use	Spanish	16.2 (14.9)
	English	26.9 (18.8)
Language Proficiency	Spanish	44.8 (8.2)
	English	51.4 (5.0)

Table 1: Participants' linguistic backgrounds.

⁷ 83 were born in the US; 9 and 18 participants immigrated to the US before and after the age of eight, respectively.

⁸ Among the total of 110 participants included in the study, 96 participants had an AoA of 8 or below, and the remaining 14 had an AoA of 9 or above.

3.2. Materials

Following and adapting the methodologies used in previous research, the study included a training phase and a test phase. The training phase was implemented via a cross-modal lexical decision task. During the training phase, participants listened to words, which were also shown on the screen one by one, in the appropriate orthography, and decided, on each trial, whether the word they heard was real or not, as quickly as possible. The goal of the lexical decision task was to induce phonetic recalibration by exposing participants to a sound which was acoustically ambiguous between [p] and [k]. Embedding these sounds into real words of English or Spanish biased participants to categorize them as certain phonemes despite their ambiguous acoustics.

The inclusion of orthographic forms in the training phase constitutes a deviation from the previous research tradition. It was implemented to enhance the recalibration effect by leveraging top-down influences on speech perception. Prior research has shown that priming can facilitate phonetic adaptation and word recognition (e.g., semantic priming, Meyer & Schvaneveldt, 1971; subtitles, Mitterer & McQueen, 2009; phoneme restoration effect, Warren, 1970). For instance, Keetels et al. (2016) showed that nonword orthographic information (e.g., showing “VdA” for priming /d/ responses in /a?a/ or “VbA” for inducing /b/) can also readjust a perceptual boundary between the two sounds. A recent study demonstrated that displaying the orthographic representations of stimuli containing an ambiguous sound facilitated a recalibration effect (see Caplan et al., 2021). In our study, the visual representation of the word forms served as an additional top-down cue, increasing the possibility that the ambiguous sound is perceived as a phoneme that completes a real word (e.g., /?/ is heard as /k/ in *relocate* but as /p/ in *disrepair*). While this approach differs from previous recalibration studies that used auditory presentation only, it aligns with broader findings on top-down processing in lexically guided speech perception (Ganong, 1980; Meyer & Schvaneveldt, 1971; Warren, 1970).

The lexical decision task included 18 critical stimuli with sounds intermediate between [p] and [k] in English and 18 analogous critical stimuli in Spanish. These ambiguous stimuli were designed to meet the following requirements in order to maximize the likelihood of recalibration (Caudrelier et al., 2023; Reinisch et al., 2013). First, none of the critical stimuli were English-Spanish cognates. Second, the ambiguous sound occurred in a stressed syllable as late in the word as possible (second or third syllable) (e.g., *disappear*). Third, we strived to avoid /p/ and /k/ and their voiced counterparts elsewhere in the critical stimuli and in fillers, although it proved difficult to exclude them completely. It was also impossible to exclude highly frequent voiceless and voiced alveolar stops /t/ and /d/. The lexical decision task also included 18 words with unambiguous contrasting sounds matching with the critical stimuli in length and frequency, 36 real-word fillers, and 72 nonce-word fillers generated by Wuggy (see Keuleers & Brysbaert, 2010, for the description of functions and instructions). Different L1 speakers of English and Spanish, both females in their 20's, recorded stimuli in the corresponding languages. The speaker

of Spanish was originally from Spain and had resided in the United States for two years and ten months for graduate studies at the time of recording. The speaker of English was born and raised in the Midwestern region of the United States and was also a graduate student. Recording was conducted in a sound-attenuated booth using a cardioid condenser microphone with a sampling rate of 44.1 kHz and a quantization rate of 16 bits.

To create a sound ambiguous between [p] and [k], a custom Python (version 3.10) algorithm for direct interpolation was employed, using NumPy and SciPy for numerical operations and audio processing, respectively. Specifically, the syllable-initial stop portions (burst and aspiration) in pairs of recordings of [p] and [k] words (e.g., *disrepair* and *relocate*) were blended into each other, creating 10 progressive steps with the first and the tenth steps being the original [p] and [k] and the intermediate steps progressing gradually from /p/ to /k/. The processed audio signals representing the stop portion were then added to the corresponding word frames, creating a [p] to [k] word continuum (i.e., *disappear–disakkear*).

Prior research showed that stop bursts contain information that can cue the following vowel (e.g., Liberman et al., 1967). Therefore, efforts were made to match vowels as closely as possible during the splicing procedure; however, perfect matching was not always possible (e.g., *disrepair* and *relocate*). While imperfect vowel matching could have introduced some unintended recalibration effects during training, a series of informal perceptual checks confirmed that the continua sounded natural, with listeners consistently perceiving a smooth transition from /p/ to /k/ without noticeable influence from vowel mismatches. If mismatched vowel cues had been the primary driver of stop perception, such a smooth transition would not have occurred. Instead, perception would have been disrupted, as the same burst can be heard as /p/ before /i/ but as /k/ before /a/ (Liberman et al., 1967).

The intermediate sound files (either the fifth or sixth on the continuum) were selected for exposure to an ambiguous sound during the lexical decision task. Since the continuum endpoints were natural recordings of /p/ and /k/ words rather than synthesized sounds, the midpoint was expected to reflect perceptual ambiguity between the two categories. While a separate norming test was not conducted, informal pilot tests demonstrated ambiguity at the intermediate steps.

The test phase consisted of a 2AFC task. The task involved 80 critical stimuli, resulting from the four 10-step continua based on /p/-/k/ minimal pairs in both English and Spanish, such as *cat-pat*, with critical segments in the word-initial position. The English minimal pairs were monosyllabic, and the Spanish minimal pairs were disyllabic with stress falling on the first syllable. The members of each minimal pair were matched in lexical frequency to avoid a frequency-based Ganong effect (Connine et al., 1993; Ganong, 1980). The intermediate steps in the continua were created by using the same method that was used to create ambiguous training stimuli for the lexical decision task. Minimal pairs were used to create the continua for the test phase, so the following vowels were always matched in the spliced stops. It should be noted that

the end points of the continua used in the 2AFC task were natural productions of /p/ and /k/, which is different from previous studies where (re)synthesized stimuli were used as endpoints (e.g., Caudrelier et al., 2024; Reinisch et al., 2013). The minimal pairs used in the task are shown in **Table 2**. The full lists of training stimuli are provided in Appendix A.

English stimuli	Spanish stimuli	
Minimal pair [p] - [k]	Minimal pair [p] - [k]	Gloss
pat-cat	paso-caso	step-case
pin-kin	peso-queso	weight-cheese
pan-can	para-cara	stop (verb, imperative)-face
pod-cod	pasa-casa	raisin-house

Table 2: The list of minimal pairs used as critical stimuli in the 2AFC task.

3.3. Procedures

All participants signed an electronic consent form and performed a headphone screening task before starting the experiment. The instruction language was always English in both the training and test tasks. Appendix B provides sample task images.

3.3.1. Lexical decision task (training)

After the screening, participants were randomly assigned to one of the four groups: English [p] ambiguous, English [k] ambiguous, Spanish [p] ambiguous, and Spanish [k] ambiguous. The BLP scores were similar across the four groups, ranging from 32.4 to 37.4. Each group designation represents the training language and the ambiguous sound to which participants were exposed during the lexical decision task. For instance, a participant belonging to the English [p] ambiguous group listened to 18 English words with an ambiguous [p] sound. Before starting the task, participants were informed that they were going to hear either English or Spanish words (depending on the group assignment). They were asked to determine whether each word was real or not by mouse-clicking on a button. They were instructed to respond as quickly as possible and informed that the trial would automatically advance if no response was provided within three seconds. On each trial, participants clicked ‘Yes’ for real words or ‘No’ for nonce words. Each word stayed on the screen for 500 ms in English or Spanish orthography (depending on the group) below a center fixation cross, to maximize the likelihood of recalibration (Samuel, 2001). Upon the completion of each stimulus playback, two response buttons appeared in the center of the screen. The ‘Yes’ and ‘No’ buttons were displayed in rectangular boxes, and their locations were counterbalanced across trials (left or right). The task included a total of 144 trials: 18 critical stimuli, 18 words with unambiguous contrasting sounds, 36 fillers, and 72 nonce words.

3.3.2. 2AFC identification task (test)

Upon completion of the lexical decision task, participants proceeded to the 2AFC task. They listened to words with initial stops that fell on the acoustic continuum between /p/ and /k/ and could be identified as either sound, given that each word was a member of a minimal pair. Participants were informed that they were going to hear either Spanish or English words (depending on the current task) and asked to decide which of the two members of the minimal pair they heard on each trial as quickly as possible (e.g., *cat* or *pat*). Each trial was preceded by 500 ms of a blank screen with a fixation cross in the center, and the response options appeared after the playback of each stimulus. Participants registered their responses with a mouse click on one of the response buttons, which were shown in the center of the screen, their order counterbalanced across participants. Upon response or after 3,000 ms had elapsed, the experiment automatically moved to the next trial. The task included a total of 240 trials (4 minimal pairs \times 2 languages \times 10 steps \times 3 repetitions). The order of languages (English first or Spanish first) was counterbalanced across participants. After the 2AFC task, participants completed the modified BLP questionnaire. The whole experiment took 30 minutes to complete, on average.

3.4. Analyses

3.4.1. Lexical decision task (training)

Participants' responses in the lexical decision task ('yes' or 'no'), a total of 15,949 data points were submitted to a Bayesian mixed-effects logistic regression model for statistical analysis, with a binary dependent variable of accuracy (1: correct, 0: incorrect). The population (fixed) effects of the model included Language (English or Spanish), BLP dominance score, Stimulus Type (Critical words, Filler nonce words, or Filler real words), the interaction between Language and BLP, and the interaction between Language and Stimulus Type, while including BLP as random intercepts and slopes for both item and subject. The categorical factors of Language and Stimulus Type were sum-coded. Since BLP was a continuous variable, the model interpreted its effects relative to a value of zero, which represents a perfectly balanced bilingual. Weakly informative priors were integrated into the model with the mean value of 0 and a standard deviation (*SD*) of 10.

3.4.2. 2AFC identification task (test)

A total of 25,713 data points were submitted to statistical analyses, excluding 1,041 data points that were 2 *SDs* above or below the group average reaction time (*RT*). *RT* was measured from the moment response buttons appeared on the screen to the moment participants registered a response on each trial, following the approach in previous studies (Johnson & Babel, 2010; Stevenson, 1973; Wrembel et al., 2019, among others).

Two Bayesian mixed-effects logistic regression models were implemented. We fitted separate models for each test language because this approach provided better predictive accuracy and

interpretability of the results than a full model incorporating both test languages, and it allowed for isolating the effects of training language on recalibration, following Caudrelier et al. (2023). The models examined the effects of training language and training category on the categorization decision ([p] or [k], or odds ratio of /k/ responses) in each of the two test languages separately. The models incorporated fixed effects of Training Language (Spanish, English), Training Category (ambiguous /p/ or ambiguous /k/), Step (10 steps of the continuum), and their two- and three-way interactions, as well as by-subject and by-item intercepts and slopes for Step as random effects. Step was centered such that a value of zero corresponded to the mean of the continuum. Categorical factors were sum-coded for better interpretability, since all categorical factors comprised two levels. Therefore, the resulting intercepts equal the grand means of the means of all levels of the categorical factors at the midpoint of the step continuum, with the beta coefficients corresponding to the deviations (main effects) of a level of a factor from the intercept, averaged for all other categorical factors at the midpoint of the step continuum.

All statistical analyses were performed using R version 4.2.1 (R Core Team, 2022) and the ‘brm’ package (Bürkner, 2017) to fit the Bayesian models via the Hamiltonian Markov Chain Monte Carlo (MCMC) sampler function in Stan (Carpenter et al., 2017). For each model, 4 MCMC chains generated 4,000 samples, following 1,000 warm-up iterations to obtain posterior samples. The present study reports the summaries of posterior distributions, including 95% credible intervals (CIs) for each point estimate. To assess the credibility of effects, Probability of Direction (*pd*), also referred to as *Maximum Probability of Effect*, was computed, which is also provided where applicable. The index, *pd*, refers to the probability that the effect of a parameter is skewed in a certain direction, be it positive or negative. A *pd* of 50% indicates a posterior of zero, suggesting no evidence for an effect on the dependent variable while a *pd* close to 100% supports evidence for an effect with either positively or negatively skewed distributions. For instance, if 95% of the posterior distribution of a certain effect is skewed away from 0 in either a negative or positive direction, it is inferred that the effect indeed exists with a probability of 95%. Due to the complexity of the influence of priors on the posterior distribution, there is currently no consensus regarding the most credible index of the effect, and *pd* is known to provide an intuitively interpretable measure comparable to the *p* value in the Frequentist framework (see Makowski et al., 2019a, 2019b for more discussion).

3.5. Theory-specific predictions

Returning to the representational assumptions of the three models of L2 speech learning, SLM-r posits that L1 and L2 speech sounds interact at the level of context-specific allophones, thus forming composite categories and affecting each other exclusively at this level. Given that our training stimuli were intersonorant, word-medial realizations of voiceless stops, place of articulation boundary recalibration induced for these sounds should be limited to this specific position across

languages. Therefore, by extension, SLM-r does not predict generalization of recalibration both across languages and across positions (although generalization across positions within language is not precluded). Since our test words contained critical segments in the word-initial position, we would expect to see recalibration within language but not across languages. PAM-L2, in contrast, assumes that L2 sounds assimilate to L1 phonemes, not only to corresponding position-specific L1 allophones. Therefore, according to PAM-L2, cross-linguistic links could transcend the boundaries of specific allophones and extend to the phoneme as a whole (including other allophones). As a result, we would expect to see recalibration trained on word-medial stops in language 1 generalize to word-initial stops in language 2. Finally, L2LP, which assumes separate L1 and L2 grammars, would not predict cross-linguistic generalization of recalibration under any condition (assuming that such generalization indicates that L1 and L2 categories interact in a shared phonetic space).

4. Results

4.1. Lexical decision accuracy (training)

The primary goal of the lexical decision task was to induce recalibration by exposing participants to ambiguous /p/ or /k/. Therefore, the present study does not focus on the examination of participants' performance in the lexical decision task, although we provide the full statistical model results for the lexical decision task in **Table 3**. Descriptive statistical analysis, complemented with a Bayesian independent *t*-test, indicated that participants were more accurate in distinguishing real words from nonce words in English (percent correct = 90%) than in Spanish (percent correct = 85%) ($BF_{10} = 21.1$). **Figure 2** visually confirms this result with the distribution of English data being skewed towards higher scores than that of Spanish data. The results from the Bayesian mixed-effects logistic regression model also indicated that participants were more accurate in English than in Spanish ($\beta = -0.47$, 95% CI = $[-1.05, 0.09]$, $pd = 94.80$), suggesting the possibility that they were somewhat more proficient in English than in Spanish.

The mean proportion correct for English critical words was 0.92 ($SD = 0.27$), while for Spanish critical words, it was 0.72 ($SD = 0.45$). Every English critical stimulus had an accuracy rate above chance (minimum = 0.65), whereas 10 out of 36 Spanish critical stimuli had a proportion correct of 0.50 or lower. This discrepancy suggests that some Spanish critical words may have been harder to recognize, presumably due to lower Spanish proficiency of our participants. At the participant level, Spanish word recognition accuracy ($M = 0.72$, $SD = 0.41$) was generally lower and more variable than in English ($M = 0.92$, $SD = 0.21$). Among the participants tested in Spanish, 10% had a mean accuracy below 0.50 on critical items. In contrast, the minimum accuracy in the English condition was 0.65. On the other hand, 91% of the participants in the English condition had an accuracy rate above 0.80, with many reaching perfect accuracy. This discrepancy suggests that for a subset of participants, Spanish critical words were not consistently recognized as real words, which may have influenced the strength of recalibration effects in the Spanish condition.

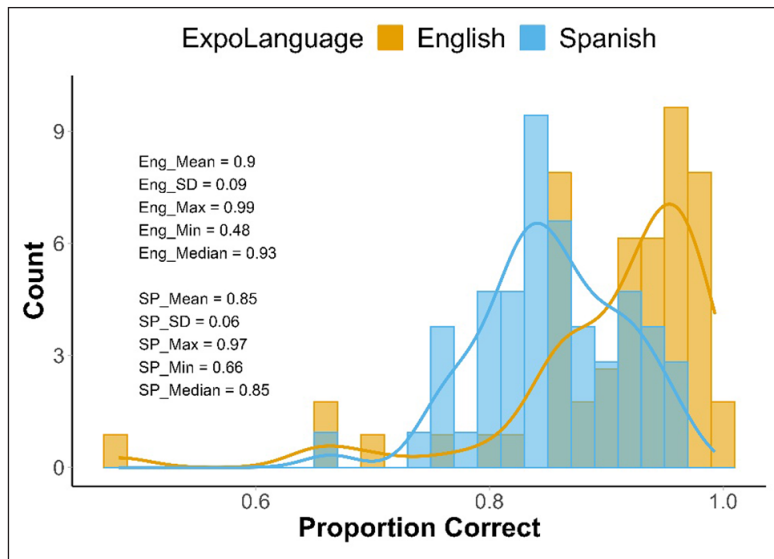


Figure 2: Histogram of participants' proportion correct lexical decision by training language. ExpoLanguage refers to the training/exposure language.

Group-level effects	Estimate	Est.Error	L-95% CI	U-95% CI	Rhat	
~Subject						
sd(Intercept)	0.92	0.08	0.77	1.10	1.00	
sd(BLP)	0.00	0.00	0.00	0.01	1.03	
cor(Intercept,BLP)	-0.10	0.52	-0.94	0.87	1.00	
~Word						
sd(Intercept)	1.63	0.10	1.46	1.83	1.01	
sd(BLP)	0.00	0.00	0.00	0.01	1.01	
cor(Intercept,BLP)	-0.30	0.30	-0.92	0.28	1.00	
Population-level effects	Estimate	Est.Error	L-95% CI	U-95% CI	Rhat	pd (%)
Intercept	2.89	0.17	2.54	3.23	1.01	100
English vs. Spanish	0.76	0.18	0.41	1.11	1.01	100
BLP	0.00	0.00	0.00	0.01	1.01	85.55
Critical vs. FillerReal	-0.09	0.10	-0.28	0.11	1.00	80.83
FillerNonce vs. FillerReal	0.37	0.16	0.06	0.68	1.00	99.15
English vs. Spanish:BLP	-0.01	0.00	-0.01	0.00	1.00	97.78
English vs. Spanish:Critical vs. FillerReal	0.42	0.10	0.22	0.61	1.01	100
English vs. Spanish:FillerNonce vs. FillerReal	-0.84	0.16	-1.15	-0.53	1.02	100

Table 3: Summary of the Bayesian mixed-effects logistic regression model examining lexical accuracy.

4.2. 2AFC task (test)

This section presents descriptive statistics from the 2AFC test phase. The mean proportions of /k/ responses were calculated across a 10-step continuum for each condition, defined by Test Language, Training Language, and Training Category (Table 4). In the Spanish test, both the English and Spanish training groups showed recalibration effects as participants exposed to /k/ tokens produced more /k/ responses than those exposed to /p/ tokens. The recalibration effect was stronger in the Spanish-trained group ($M = .65$ for /k/ training vs. $M = .41$ for /p/ training) than in the English-trained group ($M = .58$ for /k/ training vs. $M = .53$ for /p/ training). On the other hand, the recalibration effect manifested only in the test-training-matched condition in the English test. Specifically, while participants trained on English /k/ produced more /k/ responses than those trained on English /p/ ($M = .49$ for /k/ training vs. $M = .40$ for /p/ training), no meaningful difference was observed between Spanish /k-/p/ training groups ($M = .48$ for /k/ training vs. $M = .49$ for /p/ training). These descriptive patterns are consistent with the model estimates reported in the next section.

Test Language	Training Language	Training Category	Mean proportion of /k/ (SD)
Spanish	English	/p/	0.53 (0.49)
	English	/k/	0.58 (0.50)
	Spanish	/p/	0.41 (0.50)
	Spanish	/k/	0.65 (0.50)
English	English	/p/	0.40 (0.50)
	English	/k/	0.49 (0.49)
	Spanish	/p/	0.49 (0.49)
	Spanish	/k/	0.48 (0.48)

Table 4: Summary of the proportions of /k/ responses by Test Language, Training Language, and Training Category, averaged for steps.

4.3. Effects of training language on phonetic recalibration by test language

A Bayesian mixed-effects logistic regression model was implemented for each test language. Each model examined the effects of Training Category, Step (centered), Training Language and their two- and three-way interactions on the log-odds of /k/ responses. Since Step is centered, the effect of either Training Category or Training Language assumes Step at zero, which is the midpoint of the continuum. That is, the main effects of the categorical factors indicate changes in the log-odds of /k/ responses when auditory stimuli are the most ambiguous. In the subsequent sections, particular focus is placed on the following three effects in examining recalibration: (1) Step, (2) Training Category \times Step, and (3) Training Category \times Step \times Training Language. Tables 5 and 6 report all statistical results from each model.

4.3.1. Spanish test

In the Spanish test, the main effect of Training Category had a credible effect on the log-odds of /k/ response at the midpoint of the stimuli continuum, when averaged for Training Language ($\beta = -1.01$, 95% CI = $[-1.38, -0.67]$, $pd = 100.00$). This suggests that participants trained to expand the /p/ category were more likely to perceive the intermediate step as /p/ than those trained on /k/. The model predicted a substantial increase in the odds ratio of /k/ response by a factor of 3.71 per unit increase in Step, averaged across other factors ($\beta = 1.31$, 95% CI = $[0.89, 1.71]$, $pd = 99.98$). This result confirms the effectiveness of the audio stimuli by demonstrating that the resynthesized Spanish stimuli formed a /p-/k/ perceptual continuum.

The interaction between Category and Step showed a credible effect by reducing the odds of /k/ response by a factor of 0.80 per step in the /p/ group, relative to the /k/ group, averaged for Training Language ($\beta = -0.22$, 95% CI = $[-0.32, -0.12]$, $pd = 100.00$). This result, along with the main effect of Training Category, confirms that phonetic recalibration occurred, with /k/ group participants being more likely to perceive ambiguous stimuli as /k/ than /p/ group participants, as expected, when averaged for Training Language. The interaction effect of Training Category with Training Language was also credible, suggesting that the difference in the odds of /k/ response between /p/ and /k/ training groups varied by their training language ($\beta = 0.64$, 95% CI = $[0.29, 0.99]$, $pd = 100.00$). This result suggests that recalibration was more robust in the Spanish training group than in the English training group. Likewise, the three-way interaction of Training Category with Step and Training Language showed a credible effect on the odds ratio of /k/ response with a positive coefficient value ($\beta = 0.10$, 95% CI = $[0.00, 0.20]$, $pd = 97.47$). Specifically, the odds of /k/ response increased by a factor of 1.11 in the English /p/ training group with each subsequent step, compared to the increase in the Spanish /p/ training group, decreasing the gap in the probability of /k/ response between /p/ and /k/ training groups. In other words, participants who had been exposed to English /p/ were more likely to perceive stimuli as /k/

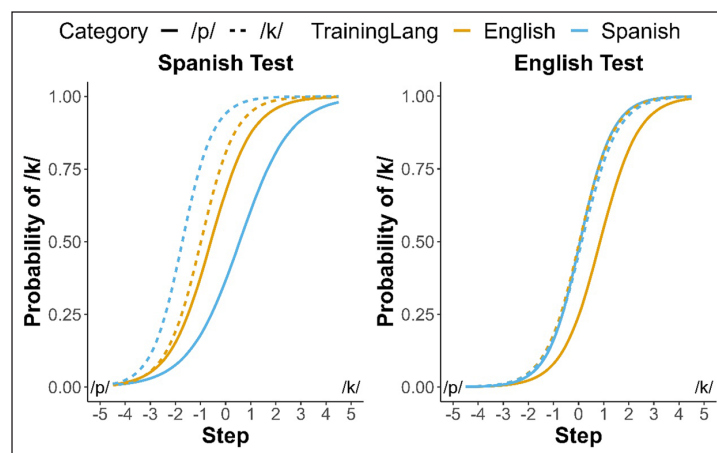


Figure 3: Bayesian logistic curves representing the predicted probability of /k/ responses across a 10-step centered continuum for Spanish (left panel) and English (right panel) posttests based on the conditional effects (Training Language, Training Category, Step).

than those who had been exposed to Spanish /p/, suggesting a mitigated recalibration effect in the English training group. On the other hand, this suggests a stronger recalibration effect in the Spanish training language. Nevertheless, the magnitude of this three-way interaction ($\beta = 0.10$) was not strong enough to offset the recalibration effect in the English training group, suggesting that a significant recalibration effect manifested itself not only in the Spanish training group but also in the English training group. The recalibration effects in the Spanish test are visualized in the left panel of **Figure 3**, where the /k/ and /p/ categorization curves are offset by a smaller distance in the English training group than in the Spanish training group. **Table 5** provides a summary of the Bayesian mixed-effects logistic regression model examining the Spanish posttest data.

Group-level effects	Estimate	Est.Error	L-95% CI	U-95% CI	Rhat	
~Subject						
sd(Intercept)	1.83	0.14	1.58	2.13	1.01	
sd(Step)	0.47	0.04	0.39	0.57	1.00	
cor(Intercept,Step)	0.69	0.06	0.56	0.80	1.00	
~Pair						
sd(Intercept)	2.14	0.97	0.95	4.79	1.00	
sd(Step)	0.30	0.24	0.09	1.02	1.01	
cor(Intercept,Step)	0.13	0.45	-0.74	0.85	1.00	
Population-level effects	Estimate	Est.Error	L-95% CI	U-95% CI	Rhat	pd (%)
Intercept	1.08	1.15	-1.34	3.56	1.00	86.12
Category/p/vs./k/	-1.01	0.18	-1.38	-0.67	1.01	100
Step	1.31	0.21	0.89	1.71	1.02	99.98
TrainLanguageEnglishvs. Spanish	-0.02	0.17	-0.37	0.32	1.01	55.25
Category/p/vs./k/:Step	-0.22	0.05	-0.32	-0.12	1.01	100
Category/p/vs./k/:TrainLan- guageEnglishvs.Spanish	0.64	0.18	0.29	0.99	1.01	100
Step:TrainLanguageEng- lishvs.Spanish	0.01	0.05	-0.09	0.11	1.00	59.80
Category/p/vs./k/:Step:Train- LanguageEnglishvs.Spanish	0.10	0.05	0.00	0.20	1.01	97.47

Table 5: Summary of the Bayesian mixed-effects logistic regression model examining the effects of training language on recalibration in Spanish.

The dependent variable is the log-odds of /k/ response. The credible negative coefficient for Training Category indicates that participants exposed to ambiguous /k/ stimuli were more likely to respond /k/ than those exposed to ambiguous /p/, averaged across Training Language and with Step centered at zero (i.e., at the ambiguous midpoint of the continuum). The credible positive coefficient for Step indicates that the odds ratio of /k/ responses increases as stimuli move toward the /k/ end of the continuum.

4.3.2. English test

Like the Spanish model, the statistical model for the English 2AFC data showed a credible effect of Training Category at the midpoint of the stimuli continuum, averaged for Training Language ($\beta = -0.25$, 95% CI = $[-0.48, -0.03]$, $pd = 98.55$). There was a credible and strong effect of Step on the odds ratio of /k/ response, with the odds increasing by a factor of 4.22 for each unit increase in Step, averaged for other factors ($\beta = 1.44$, 95% CI = $[1.02, 1.88]$, $pd = 100.00$). This indicates that participants were increasingly likely to perceive stimuli as /k/ words as the step increased, confirming the effectiveness of the English audio stimuli. Unlike the Spanish model, the effect of Training Language on the log-odds of /k/ response was credible at the midpoint of the continuum, averaged for Training Category ($\beta = -0.22$, 95% CI = $[-0.46, 0.01]$, $pd = 96.95$). This suggests that participants trained in English were less likely to perceive stimuli as /k/ in the English test than those trained in Spanish, when averaged for the category. However, this effect is due to the absence of the recalibration effect in the Spanish-training group.

The interaction effect of Training Category with Training Language was credible at the midpoint of the continuum ($\beta = -0.29$, 95% CI = $[-0.52, -0.06]$, $pd = 99.42$), indicating that the odds of /k/ response in the English /p/ training group were significantly lower compared to in the Spanish /p/ training group, when auditory stimuli were ambiguous. This interaction effect suggests a more robust recalibration effect in the English training group than in the Spanish training group. The three-way interaction among Category, Step, and Training Language explains that the likely reason for the observed effect of Training Language at the midpoint step, averaged for Training Category, was the absence of recalibration in the Spanish training group, as suggested by the negative coefficient of the three-way interaction ($\beta = -0.07$, 95% CI = $[-0.15, 0.01]$, $pd = 95.35$). The increase in log-odds with incremental steps in the /p/ training group was more pronounced in the Spanish training group than in the English training group. This suggests that the difference in the odds of /k/ response between the /p/ and /k/ groups along the /p-/k/ continuum was greater for participants exposed to English, indicating a more robust recalibration effect. In contrast to the Spanish test, the magnitude of the three-way interaction ($\beta = -0.07$) was strong enough to offset the recalibration effect in Spanish, indicating that recalibration was present in the English training group but not in the Spanish training group, in the English test. The absence of recalibration in the Spanish training group is visually confirmed in the right panel of **Figure 3**, with the logistic curves of /k/ and /p/ training categories fully overlapping in the Spanish training group.

To summarize, different perceptual patterns were observed in participants' categorization of /p-/k/ stimuli across different test languages. The common finding is that phonetic recalibration was always evident when the test language matched the training language. In the Spanish test, strong evidence of recalibration was found in the Spanish training group. Similarly, recalibration occurred in the English test for the English training group. However, asymmetric perceptual

patterns were observed in the two cases of unmatched training and test language. Recalibration was present in the English training – Spanish test group, although the degree of recalibration was significantly smaller than in the Spanish training – Spanish test group. In contrast, no evidence of recalibration was found for the Spanish training – English test group.

Group-level effects	Estimate	Est.Error	L-95% CI	U-95% CI	Rhat	
~Subject						
sd(Intercept)	1.15	0.09	0.98	1.35	1.00	
sd(Step)	0.36	0.04	0.30	0.44	1.00	
cor(Intercept,Step)	0.04	0.12	-0.20	0.27	1.00	
~Pair						
sd(Intercept)	2.48	1.13	1.14	5.47	1.00	
sd(Step)	0.37	0.27	0.12	1.11	1.00	
cor(Intercept,Step)	0.26	0.42	-0.63	0.88	1.00	
Population-level effects	Estimate	Est.Error	L-95% CI	U-95% CI	Rhat	pd (%)
Intercept	-0.37	1.27	-2.96	2.34	1.00	63.95
Category/p/vs./k/	-0.25	0.12	-0.48	-0.03	1.01	98.55
Step	1.44	0.20	1.02	1.88	1.00	100
TrainLanguageEnglishvs. Spanish	-0.22	0.12	-0.46	0.01	1.00	96.95
Category/p/vs./k/:Step	-0.01	0.04	-0.09	0.07	1.00	57.12
Category/p/vs./k/:TrainLan- guageEnglishvs.Spanish	-0.29	0.12	-0.52	-0.06	1.00	99.42
Step:TrainLanguageEng- lishvs.Spanish	-0.04	0.04	-0.12	0.03	1.00	87.12
Category/p/vs./k/:Step:Train- LanguageEnglishvs.Spanish	-0.07	0.04	-0.15	0.01	1.00	95.35

Table 6: Summary of the Bayesian mixed-effects logistic regression model examining the effects of training language on recalibration in English.

The dependent variable is the log-odds of /k/ response. The credible negative coefficient for Training Language indicates that participants trained in Spanish were more likely to respond /k/ than those trained in English, averaged across Training Category and with Step centered at zero. The interpretations of Training Category and Step follow the same principles as described in Table 5.

4.4. Effects of bilingual dominance on phonetic recalibration

As suggested by an anonymous reviewer, we conducted additional analyses to examine the potential effects of bilingual dominance (as measured by BLP) on phonetic recalibration. We

fit two separate Bayesian linear regression models for the Spanish and English tests, using participants' proportion of /k/ responses at the midpoint of the continuum (Steps 5 and 6), where recalibration was most prominent, as the dependent variable. The models included Training Category, Training Language, and BLP score, along with their interactions. The categorical factors were sum-coded. However, while both models indicated credible effects of Training Language and the interaction between Category and Training Language, as observed in prior models, none of the BLP or BLP interaction effects were found credible ($pd < 95\%$).

5. Discussion

The current study investigated the possibility of cross-linguistic transfer of the effects of lexically guided phonetic recalibration from one language to another in bilingual speakers. Specifically, our primary question was whether recalibration trained in one language (e.g., English), can manifest in the other language, Spanish, with listeners who are Spanish-English bilinguals. Evidence of such transfer would be consistent with the hypothesis that corresponding sound categories are not fully independent across the two languages and instead are integrated or linked to each other at some level of representation. If proven effective, this methodology can be used for investigating cross-linguistic organization of sound categories in language learners and bilinguals in conjunction with other methodologies, such as acoustic studies of cross-linguistic influence between L1 and L2 categories.

5.1. Shared or separate categories?

Regarding the question of shared or separate L1 and L2 categories in bilinguals, previous work examining bilingual phonetic recalibration agrees that in cases of high phonetic similarity between L1 and L2 sounds, as exemplified by /s/ and /f/ across languages such as Dutch, German, English, and French, these sound categories could be shared across languages, given robust evidence of cross-linguistic transfer of recalibration (Caudrelier et al., 2023; Reinisch et al., 2013; Schuhmann, 2016). We modified several key methodological parameters in comparison to those in the previous work, which allowed us to formulate a related but novel research question: Does transfer of recalibration occur with sound categories that are phonetically different but phonologically related across languages, such as Spanish voiceless unaspirated and English voiceless aspirated stops, and in the absence of the possible facilitative effects of speaker-specific learning? Concomitantly, we tested generalization of recalibration across phonetic contexts (from word-medial to word-initial), which allowed us to probe the question of the level of representation at which cross-category links are established between languages. To answer these questions, our participants learned to retune the boundaries of the place of articulation contrast between word-medial voiceless stops /p/ and /k/ in English and Spanish through a cross-modal lexical decision task, and while half of them were tested on the same language and the same

speaker in a 2AFC task, the second half were tested on a different language and a different speaker. The test presented critical segments in the word-initial position.

The main purpose of the lexical decision task was to induce phonetic recalibration between the /p/ and /k/ categories. Nonetheless, the statistical analysis of the data yielded findings that merit attention. Specifically, the accuracy results suggested that participants were somewhat more proficient in English than in Spanish, as accuracy was higher in English (90%) than in Spanish (85%), aligning with the BLP findings that participants were more English dominant. In addition, these accuracy rates were more consistent with advanced L2 speakers' results and were somewhat lower than monolingual speakers' accuracy (95%) reported in previous studies (Reinisch et al., 2013; Schuhmann, 2016).

The primary goal of the 2AFC task was to investigate the potential cross-linguistic transfer of phonetic recalibration, while also testing for generalization across contexts and, in some conditions, across speakers. The statistical analyses indicated both symmetric and asymmetric perceptual behaviors among the participants. First, perceptual categorization was strongly dependent on the steps of the acoustic /p/-/k/ continuum in all models, confirming that acoustic resynthesis aimed at creating progressive /p/-/k/ steps was effective. Second, perceptual recalibration of phonemic boundaries was robustly present when training language and test language (as well as speaker) matched, both for Spanish and English. This finding confirms that the training technique was effective overall. It also demonstrates successful generalization of recalibration from word-medial to word-initial contexts within languages. Additionally, this finding adds to a small body of research indicating that listeners need not be monolingual/L1 speakers of a given language for perceptual learning of this nature to take place (Caudrelier et al., 2023, 2024; Llompert, 2024; Reinisch et al., 2013; Schuhmann, 2016). Specifically, Spanish-English bilingual speakers in the current study who were mostly heritage speakers of Spanish and who differed from those investigated in previous research (diglossic French-English bilinguals in Montreal, Canada, or proficient L2 speakers of closely related languages: Dutch-English or German-English), recalibrated effectively in both of their languages. It should also be noted that in our study, when the training and test languages matched, the speaker also matched, which could have boosted perceptual learning in these conditions.

The two groups who were tested in the nonmatching condition between training and test languages demonstrated asymmetric perceptual patterns. Those who were trained on English and tested on Spanish showed evidence of recalibration generalization from English to Spanish. This finding suggests that bilinguals' Spanish and English voiceless stops share a connection across languages. Importantly, in this case, recalibration generalized both across positions and across languages. This finding is consistent with the assumption of PAM-L2 that cross-linguistic links are not limited to the level of position-specific allophones, as postulated by SLM-r, but occur at the phonemic level. For recalibration to take place, participants had to generalize from word-medial

allophones of English voiceless stops to word-initial allophones of Spanish voiceless stops, suggesting that these sounds are integrated with each other in some way, possibly as parts of a unified interlanguage phoneme. This finding is also at odds with the prediction derived from L2LP, according to our understanding of its assumptions, that cross-linguistic integration among sound categories cannot take place because the two grammars are separate and independent. Of course, these findings can support or contradict relevant theoretical assumptions only to the extent to which they can be interpreted as evidence of cross-linguistic links between L1 and L2 sounds.

Although the magnitude of the effect was decidedly smaller in the unmatched than in the matched conditions, it should be taken into account that the speaker was different between the English training and the Spanish test, possibly diminishing the effect. Recalibration of stop contrasts has also been shown to elicit weaker effects than recalibration of fricatives (Kraljic & Samuel, 2006). At the same time, matching the gender of two different speakers, and the use of stops instead of fricatives were associated with greater likelihood of cross-speaker generalization in previous work, in agreement with our findings (Kraljic & Samuel, 2005, 2006, 2007; Tamminga et al., 2020).

In contrast, the analysis revealed no evidence of recalibration transfer from Spanish to English in participants who were trained in Spanish but tested in English, suggesting that cross-linguistic generalization may not have taken place in this case. There are several possible interpretations or explanations for this asymmetry.

5.2. Possible asymmetry explanations

First, the results could be genuinely indicative of L1 and L2 sound categories that are linked across languages in an asymmetric manner, such that a more dominant language exerts a stronger influence on the less dominant one (Caramazza et al., 1973; Kartushina & Martin, 2019; Mack, 1989). In support of this hypothesis, in Caudrelier et al. (2024), cross-linguistic generalization of recalibration was stronger when exposure stimuli were in participants' dominant language. On the other hand, our exploratory model of dominance failed to support this explanation, given the lack of credible effect for the BLP dominance score. However, since BLP is a self-reported, subjective measure, it may not fully assess language dominance. Without incorporating an external, objective measure of dominance, it remains premature to entirely rule out its potential role in the observed asymmetry in recalibration.

Another possibility relates to the fact that comparable phonetic realizations of stop consonants across English and Spanish do not map onto comparable phonological voicing categories. Specifically, English word-initial *voiced* stops are typically realized as voiceless unaspirated, similarly to Spanish word-initial *voiceless* stops (Lisker & Abramson, 1964). This may have compromised the perceptual link from Spanish voiceless unaspirated stops to English voiceless aspirated stops, explaining the asymmetric pattern.

Related to this scenario is the possibility that aspirated stops provide more salient acoustic cues to the place of articulation than unaspirated stops because they are more extended in time. The more salient acoustic nature of aspirated stops may lead to more successful recalibration, including its generalizability to untrained languages (see also Wright, 2001, 2004).⁹ This possibility needs to be addressed in future research using phonemic categories that do not present such an asymmetry in terms of the robustness of perceptual cues.

The instruction language may have also influenced recalibration in the present study by activating a specific language mode (Grosjean, 2001, p. 49). Since instructions were always given in English, regardless of the training or test language, it is plausible that participants' English mode remained more strongly activated throughout the experiment, compared to Spanish, modulating recalibration effects.

Finally, it is possible that the Spanish stimuli were more effective at yielding or revealing recalibration than the English stimuli. As pointed out by an anonymous reviewer, we acknowledge that the imperfect match in CV syllables during the splicing procedure may have contributed to the observed asymmetry in recalibration. Since English voiceless stops have longer burst durations than Spanish ones, the residual vowel information in the burst may have been more pronounced in the English exposure stimuli. This could have introduced stronger coarticulatory cues, potentially influencing how ambiguous sounds were categorized and contributing to the asymmetry in recalibration patterns. Thus, a stimuli-related explanation cannot be entirely ruled out. However, it should be noted that the vowel information was unlikely to be the primary driver of perceptual patterns in the 2AFC task as this would have disrupted the observed smooth perceptual transition from /p/ to /k/ (Liberman et al., 1967).

5.3. Object of recalibration

Our findings speak to another issue, although it was not at the core of our investigation. An unresolved debate in the literature concerns the object of recalibration. That is, whether recalibration affects phonemes, phonological features, allophones, acoustics cues, or highly context-specific tokens. In other words, there remains a question regarding the abstractness versus specificity of the object of recalibration. The answer to this question is complicated by the fact that evidence supporting both sides of the spectrum has been reported in the literature (Dahan & Mead, 2010; Eisner & McQueen, 2005; Jesse & McQueen, 2011; McQueen et al., 2006; Mitterer et al., 2016; Reinisch et al., 2014). Generalization of recalibration across contexts and speakers suggests perceptual learning that abstracts away from the specifics of the training set. Thus, on the face of it, our findings appear to support the view that the object of recalibration is rather abstract, such as phoneme or a phonological feature. Our participants were trained to

⁹ We thank Dr. Jongho Jun for pointing this out in his discussion of this work at LabPhon 19.

recalibrate the place of articulation based on word-medial, inter-sonorant stops, but tested on word-initial, prevocalic stops. The fact that recalibration occurred indicates that generalization over word positions and phonetic contexts took place. Furthermore, in the English training–Spanish testing condition, recalibration was also generalized to a phonetically different stop consonant (from voiceless aspirated to voiceless unaspirated), as well as to a different speaker and a different language. This suggests an abstract unit, such as phonemes, as the object of recalibration (assuming that Spanish and English voiceless stops belong to the same phoneme, which crosses language boundaries), or possibly a phonological feature of place. The latter predicts that training to recalibrate place on stop consonants should generalize to other segments with the same place features (e.g., nasals). Disputing this prediction, Reinisch et al. (2014) and Mitterer et al. (2016) did not observe generalization of recalibration from stops to nasals along the same place dimension. Moreover, a body of relevant research findings reviewed in Reinisch et al. (2014) suggests that generalization tends to be found in cases where acoustic cues are highly consistent between the training and the testing segments, despite changes in the context and speaker. In support of this view, Mitterer et al. (2016) found that recalibration generalized from tensified stops to highly acoustically similar plain stops in Korean, but not from tensified stops to more acoustically dissimilar aspirated stops or nasals. Kraljic and Samuel (2005) also suggested that patterns of acoustic similarity between male and female voices provide basis for cross-speaker generalization of perceptual learning. Therefore, acoustic similarity of cues to place across aspirated and unaspirated stops, across medial and initial stops, and across different speakers could explain the generalizability of recalibration in our data, without presupposing, but not precluding, a highly abstract object of recalibration, such as a phoneme or similarly abstract underlying unit.

While our findings contribute to this important debate, they are not sufficient to resolve it. In answering this question, further research with a greater variety of phonological contrasts demonstrating distinct patterns of cue variability across contexts and across speakers will be critical.

To conclude, recalibration manifested in both languages of the bilinguals in the training–test match condition and asymmetrically generalized to the untrained language in one of the unmatched conditions. This behavioral asymmetry could have several explanations, including the difference in language dominance, the nature of the phonological contrast tested, or stimuli-related reasons. Thus, our findings paint a complex picture of cross-linguistic organization of sound categories, partially supporting the possibility of category mergers across languages. Another recent report further complicates the situation. Caudrelier et al. (2024) demonstrated that, in addition to generalizing across languages, recalibration can proceed in two opposing directions in the two languages of bilinguals, provided exposure in those languages supplies appropriate perceptual evidence. That is, bilinguals can learn to expand an /s/ category in French while shrinking /s/ in English, suggesting that each category is sufficiently autonomous.

Bilinguals' ability to make such divergent language-specific adjustments may contradict the idea that the manipulated categories are merged across languages. If so, it compromises the validity of cross-linguistic generalization of recalibration as evidence of such mergers, since both divergent and congruent cross-linguistic recalibration was observed for the same group of bilinguals in Caudrelier et al. (2024). Alternatively, it is possible that the merged L1-L2 categories can establish a degree of mutual autonomy, which nevertheless raises the question of what would constitute proof of true and complete cross-linguistic independence between related sound categories, if such true independence is possible. To further illuminate this complex question, more research needs to be conducted with diverse bilingual populations and with a greater variety of phonological categories. Nevertheless, our findings provide the first evidence that indicates that transfer of phonetic recalibration across the two languages of bilinguals is possible even in the absence of the facilitative effect of speaker-specific learning and with phonetically distinct but phonologically related categories. The fact that phonetic learning generalized across speakers, languages, and phonetic differences suggests some degree of integration, however asymmetric, between L1 and L2 sound categories (e.g., Spanish and English /p/), possibly along the lines of underlying phonological representations or unifying acoustic characteristics.

Appendix A. Stimuli lists for the lexical decision task

(a) Spanish stimuli

Number	Sound	Task	Words	Meaning	Matching noncewords
1	p	Training	depara	yield	setara
2	p	Training	deporte	sport	devalte
3	p	Training	depone	depose	degecte
4	p	Training	repaso	I review	deviso
5	p	Training	repiensa	rethink	tebienka
6	p	Training	supera	overcome	cuseta
7	p	Training	amapola	type of flower	acabozo
8	p	Training	atropella	run over	saprocua
9	p	Training	arepera	seller	tebienca
10	p	Training	ropita	baby clothes	narida
11	p	Training	arropado	wrapped up	aposato
12	p	Training	rapaces	predatory, greedy	tasaques
13	p	Training	reparto	distribution	deverto
14	p	Training	sopita	little soup	cobita
15	p	Training	mapache	raccoon	paviche
16	p	Training	tapita	lid	napida
17	p	Training	rapado	shaven	manaso
18	p	Training	apodo	nickname	apuda
19	k	Training	tacaño	mean	taripo
20	k	Training	alicate	pliers	conisate
21	k	Training	alicante	city's name	amedinte
22	k	Training	acaso	perhaps	arico
23	k	Training	acoso	harassment	aroca
24	k	Training	ocaso	twilight	orico
25	k	Training	tacada	stroke	barida
26	k	Training	requema	burn	resepa
27	k	Training	decano	dean	rerino
28	k	Training	secado	to dry (participle)	mesido
29	k	Training	recodo	a tiny small place	deroda
30	k	Training	recado	errand	desido
31	k	Training	sacude	shakes	nacude

(Contd.)

Number	Sound	Task	Words	Meaning	Matching noncewords
32	k	Training	recorta	trim	beporsa
33	k	Training	maqueta	model	pafiata
34	k	Training	recoge	pick up	seboque
35	k	Training	tocado	touch	borida
36	k	Training	sacado	taken out	namapo
37	None	Filler	caliente	hot	cacueste
38	None	Filler	señorita	miss	perosita
39	None	Filler	distante	distant	sostente
40	None	Filler	idioma	language	ipieza
41	None	Filler	ciudad	town	ciucad
42	None	Filler	maravilla	wonder	pasabibla
43	None	Filler	comprobar	check	coscrovar
44	None	Filler	completo	full	casfreto
45	None	Filler	camion	trunk	sapicon
46	None	Filler	profunda	deep	proverda
47	None	Filler	pulgadas	inch	belpasas
48	None	Filler	pregunta	question	trevinta
49	None	Filler	cuidado	care	cuebido
50	None	Filler	telefono	phono	senesopa
51	None	Filler	ventana	window	fantina
52	None	Filler	izquierda	left	isuienda
53	None	Filler	respuesta	answer	descuista
54	None	Filler	escuela	school	espiosa
55	None	Filler	deberia	should	sequesia
56	None	Filler	cubierta	cover	cupuelta
57	None	Filler	alimentos	food	apibancos
58	None	Filler	edificio	building	epevinio
59	None	Filler	primavera	spring	fricabero
60	None	Filler	aumentar	raise	ainantar
61	None	Filler	discurso	speech	soscurno
62	None	Filler	producirse	occur	rofuniuno
63	None	Filler	temporada	season	bosponada
64	None	Filler	bastante	pretty	viscante
65	None	Filler	tercero	third	tasmero

(Contd.)

Number	Sound	Task	Words	Meaning	Matching noncewords
66	None	Filler	delgado	thin	degnido
67	None	Filler	planeta	planet	clavena
68	None	Filler	movimiento	motion	sobediento
69	None	Filler	camino	path	casira
70	None	Filler	habilidad	skill	bagenidad
71	None	Filler	manzana	apple	paldana
72	None	Filler	asiento	seat	acausto

(b) English stimuli

Num	Sound	Task	Words	Matching noncewords
1	p	Training	disappear	disambiar
2	p	Training	disappoint	disapinement
3	p	Training	disrepair	disfepotle
4	p	Training	disrepute	disnevote
5	p	Training	overpay	overfying
6	p	Training	overpower	overkauer
7	p	Training	underpay	underfox
8	p	Training	underperform	underfuring
9	p	Training	underpinning	undergorring
10	p	Training	repeat	rebeet
11	p	Training	append	apite
12	p	Training	apology	aminopy
13	p	Training	unappealing	unasoling
14	p	Training	repeal	rebeel
15	p	Training	repulse	renums
16	p	Training	rapport	rappave
17	p	Training	interpersonal	indercermisal
18	p	Training	bypath	bysoth
19	k	Training	bodycam	sodyram
20	k	Training	undercover	omperbover
21	k	Training	sugarcane	sucarnats
22	k	Training	avocado	axolavi
23	k	Training	advocate	atorrate

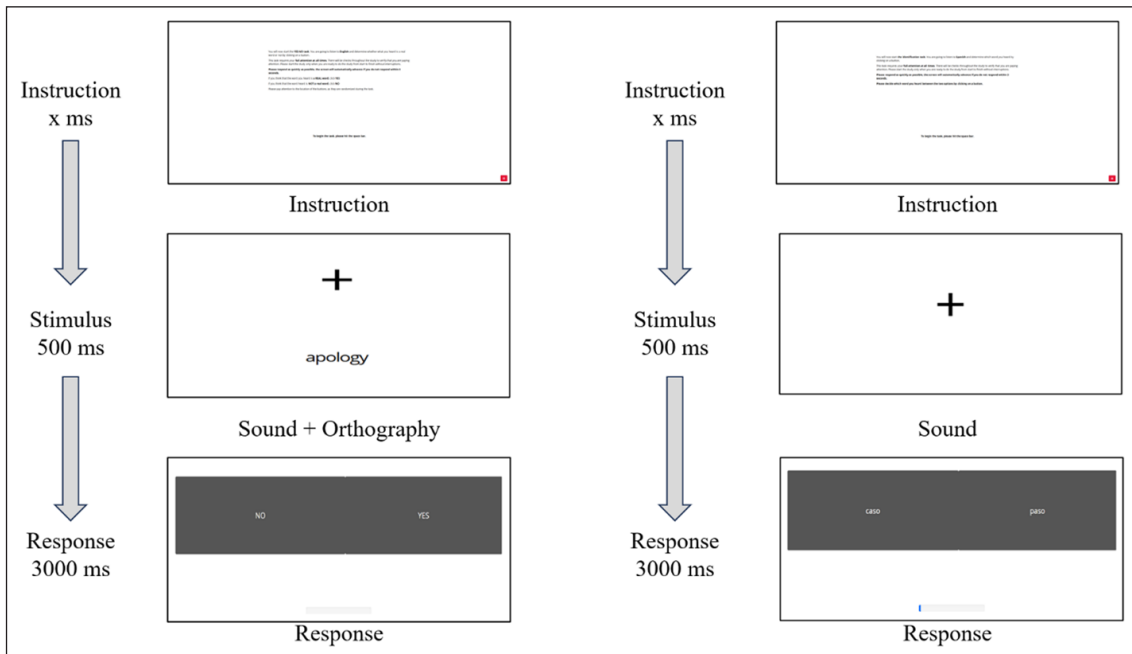
(Contd.)

Num	Sound	Task	Words	Matching noncewords
24	k	Training	allocate	attofate
25	k	Training	relocate	recofant
26	k	Training	technicality	tophrirolity
27	k	Training	recoil	resoil
28	k	Training	become	besime
29	k	Training	recook	bisool
30	k	Training	rekindle	rewintra
31	k	Training	akin	awon
32	k	Training	accountant	actooshant
33	k	Training	because	bebalts
34	k	Training	recast	reoist
35	k	Training	overcame	overtorm
36	k	Training	unaccommodating	unappommotating
37	None	Filler	agreement	acreecant
38	None	Filler	American	camarison
39	None	Filler	approach	adfloach
40	None	Filler	campaign	kimsein
41	None	Filler	community	corsarity
42	None	Filler	support	sucorp
43	None	Filler	official	occikiel
44	None	Filler	consumer	congicer
45	None	Filler	discussion	dencassion
46	None	Filler	director	dicactar
47	None	Filler	society	bemioty
48	None	Filler	situation	mesoation
49	None	Filler	experience	envenients
50	None	Filler	university	olecursity
51	None	Filler	appreciate	autholiate
52	None	Filler	conclusion	comshudion
53	None	Filler	describe	destrobe
54	None	Filler	relationship	retationtrox
55	None	Filler	fantastic	fondestic
56	None	Filler	recent	resant
57	None	Filler	recall	rebacs

(Contd.)

Num	Sound	Task	Words	Matching noncewords
58	None	Filler	applaud	appler
59	None	Filler	overload	overreen
60	None	Filler	financial	linandiel
61	None	Filler	colloquial	cospuclial
62	None	Filler	important	immestant
63	None	Filler	information	inrangation
64	None	Filler	bilingual	bilinsyas
65	None	Filler	maintain	mairbain
66	None	Filler	attention	attission
67	None	Filler	necessary	badissary
68	None	Filler	computer	cormuner
69	None	Filler	newspaper	wimbaper
70	None	Filler	operation	udalation
71	None	Filler	subscribe	substrink
72	None	Filler	decision	pevigion

Appendix B. Task procedures



Sample task screens (left: Lexical decision task, right: 2AFC task). Resolutions are adjusted.

Acknowledgements

The authors thank the two anonymous reviewers and the Guest Editor, Dr. Jeff Holliday, for their constructive and valuable feedback, which helped improve the quality of our manuscript. We also thank the Discussant, Dr. Jongho Jun, and the attendees of our presentation at LabPhon 19 for their insightful comments.

Competing interests

The authors have no competing interests to declare.

Author contributions

Yuhyeon Seo: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Olga Dmitrieva: Writing – review & editing, Writing – original draft, Project administration, Methodology, Resources, Funding acquisition, Conceptualization.

References

- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407. <https://doi.org/10.3758/s13428-019-01237-x>
- Baker, W., & Trofimovich, P. (2005). Interaction of native- and second-language vowel system (s) in early and late bilinguals. *Language and Speech*, *48*(1), 1–27. <https://doi.org/10.1177/00238309050480010101>
- Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning*, 13–34. John Benjamins Publishing Company. <https://doi.org/10.1075/llt.17.07bes>
- Birdsong, D., Gertken, L. M., & Amengual, M. (2012). Bilingual language profile: An easy-to-use instrument to assess bilingualism. *COERLL, University of Texas at Austin*.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Caplan, S., Hafri, A., & Trueswell, J. C. (2021). Now you hear me, later you don't: The immediacy of linguistic computation and the representation of speech. *Psychological Science*, *32*(3), 410–423. <https://doi.org/10.1177/0956797620968787>
- Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., & Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *The Journal of the Acoustical Society of America*, *54*(2), 421–428. <https://doi.org/10.1121/1.1913594>

- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1). <https://doi.org/10.18637/jss.v076.i01>
- Casillas, J. V. (2020). Phonetic category formation is perceptually driven during the early stages of adult L2 development. *Language and Speech*, 63(3), 550–581. <https://doi.org/10.1177/0023830919866225>
- Caudrelier, T., Martin, C. D., Samuel, A. G., Beausoleil, M. M., Tiede, M., & Ménard, L. (2023). Lexically guided phonetic recalibration transfers across languages in French-English bilinguals. In Proceedings for *the 2023 International Congress of Phonetic Sciences (ICPhS)*, 2911–2915.
- Caudrelier, T., Ménard, L., Beausoleil, M. M., Martin, C. D., & Samuel, A. G. (2024). When Jack isn't Jacques: Simultaneous opposite language-specific speech perceptual learning in French-English bilinguals. *PNAS Nexus*, 3(9), pgae354. <https://doi.org/10.1093/pnasnexus/pgae354>
- Connine, C. M., Titone, D., & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(1), 81–94. <https://doi.org/10.1037/0278-7393.19.1.81>
- Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, 34(2), 269–284. <https://doi.org/10.1016/j.wocn.2005.06.002>
- Dahan, D., & Mead, R. L. (2010). Context-conditioned generalization in adaptation to distorted speech. *Journal of Experimental Psychology: Human Perception and Performance*, 36(3), 704–728. <https://doi.org/10.1037/a0017449>
- Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, 67(2), 224–238. <https://doi.org/10.3758/BF03206487>
- Escudero, P. (2005). *Linguistic perception and second language acquisition: Explaining the attachment of optimal phonological categorization* [Doctoral dissertation, Utrecht University]. Utrecht University Repository. <https://www.google.com/url?q=https://dspace.library.uu.nl/handle/1874/7349&sa=D&source=editors&ust=1767487400856962&usg=AOvVaw0U6-myFD9I68QrOr7hk7Xn>
- Escudero, P. (2009). The linguistic perception of similar L2 sounds. In P. Boersma & S. Hamann (Eds.), *Phonology in Perception*, 15, 151–190. De Gruyter. <https://doi.org/10.1515/9783110219234.151>
- Escudero, P., & Yazawa, K. (2024). The second language linguistic perception model. In M. Amengual (Ed.), *The Cambridge handbook of bilingual phonetics and phonology* (pp. 173–195). Cambridge University Press.
- Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 92, 233–277.
- Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, 6, 319–355.
- Flege, J. E., & Bohn, O. S. (2021). The revised speech learning model (SLM-r). *Second language speech learning: Theoretical and empirical progress*, 3–83.
- Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, 15(1), 67–83. [https://doi.org/10.1016/S0095-4470\(19\)30538-8](https://doi.org/10.1016/S0095-4470(19)30538-8)

- Flege, J. E., Schirru, C., & MacKay, I. R. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication, 40*(4), 467–491. [https://doi.org/10.1016/S0167-6393\(02\)00128-0](https://doi.org/10.1016/S0167-6393(02)00128-0)
- Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*(1), 110–125. <https://doi.org/10.1037/0096-1523.6.1.110>
- Grosjean, F. (2001). The bilingual's language modes. In J. Nichole (ed.). *One mind, two languages: Bilingual language processing*, pp. 1–22. Blackwell.
- Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America, 107*(5), 2711–2724. <https://doi.org/10.1121/1.428657>
- Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review, 18*, 943–950. <https://doi.org/10.3758/s13423-011-0129-2>
- Johnson, K., & Babel, M. (2010). On the perceptual basis of distinctive features: Evidence from the perception of fricatives by Dutch and English speakers. *Journal of Phonetics, 38*(1), 127–136. <https://doi.org/10.1016/j.wocn.2009.11.001>
- Kartushina, N., & Martin, C. D. (2019). Talker and acoustic variability in learning to produce nonnative sounds: Evidence from articulatory training. *Language Learning, 69*(1), 71–105. <https://doi.org/10.1111/lang.12315>
- Keetels, M., Schakel, L., Bonte, M., & Vroomen, J. (2016). Phonetic recalibration of speech by text. *Attention, Perception, & Psychophysics, 78*, 938–945. <https://doi.org/10.3758/s13414-015-1034-y>
- Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*, 627–633. <https://doi.org/10.3758/BRM.42.3.627>
- Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141–178. <https://doi.org/10.1016/j.cogpsych.2005.05.001>
- Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262–268. <https://doi.org/10.3758/BF03193841>
- Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1–15. <https://doi.org/10.1016/j.jml.2006.07.010>
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*(6), 431–461. <https://doi.org/10.1037/h0020279>
- Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384–422. <https://doi.org/10.1080/00437956.1964.11659830>
- Llompart, M. (2024). Lexically-guided perceptual recalibration from acoustically unambiguous input in second language learners. *Journal of Phonetics, 107*, 101366. <https://doi.org/10.1016/j.wocn.2024.101366>

- Mack, M. (1989). Consonant and vowel perception and production: Early English-French bilinguals and English monolinguals. *Perception & Psychophysics*, *46*(2), 187–200. <https://doi.org/10.3758/BF03204982>
- Makowski, D., Ben-Shachar, M. S., Chen, S. A., & Lüdecke, D. (2019a). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*, 2767. <https://doi.org/10.3389/fpsyg.2019.02767>
- Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019b). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. <https://doi.org/10.21105/joss.01541>
- McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126. https://doi.org/10.1207/s15516709cog0000_79
- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227–234. <https://doi.org/10.1037/h0031564>
- Mitterer, H., Cho, T., & Kim, S. (2016). What are the letters of speech? Testing the role of phonological specification and phonetic similarity in perceptual learning. *Journal of Phonetics*, *56*, 110–123. <https://doi.org/10.1016/j.wocn.2016.03.001>
- Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS One*, *4*(11), e7785. <https://doi.org/10.1371/journal.pone.0007785>
- Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. [https://doi.org/10.1016/S0010-0285\(03\)00006-9](https://doi.org/10.1016/S0010-0285(03)00006-9)
- Olson, D. J. (2023). Measuring bilingual language dominance: An examination of the reliability of the Bilingual Language Profile. *Language Testing*, 521–547. <https://doi.org/10.1177/02655322221139162>
- R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 75–86. <https://doi.org/10.1037/a0027979>
- Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics*, *45*, 91–105. <https://doi.org/10.1016/j.wocn.2014.04.002>
- Repp, B. H. (1984). Categorical perception: Issues, methods, findings. *Speech and Language*, *10*, 243–335. <https://doi.org/10.1016/B978-0-12-608610-2.50012-1>
- Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*(4), 348–351. <https://doi.org/10.1111/1467-9280.00364>
- Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421–436. <https://doi.org/10.1006/jpho.1997.0051>

- Schuhmann, K. S. (2016). Cross-linguistic perceptual learning in advanced second language listeners. *Proceedings of the Linguistic Society of America*, 1, 31. <https://doi.org/10.3765/plsa.v1i0.3731>
- Seo, Y. (2024). *Cross-linguistic influence in L1 phonetic categories in Korean heritage speakers and long-term immigrants*. [Doctoral dissertation, Purdue University]. <https://doi.org/10.25394/PGS.25588083>
- Seo, Y., & Dmitrieva, O. (2024). L2 cross-linguistic influence on L1 perception: Evidence from heritage speakers and long-term immigrants. *Journal of Phonetics*, 104, 101314. <https://doi.org/10.1016/j.wocn.2024.101314>
- Stevenson, P. W. (1973). Reaction time measurements in speech discrimination tasks—An automated system with closed response sets. *Journal of Phonetics*, 1(4), 347–367. [https://doi.org/10.1016/S0095-4470\(19\)31403-2](https://doi.org/10.1016/S0095-4470(19)31403-2)
- Takahashi, C. (2023). L1 vowel perceptual boundary shift as a result of L2 vowel learning. *Journal of Phonetics*, 100, 101265. <https://doi.org/10.1016/j.wocn.2023.101265>
- Tamminga, M., Wilder, R., Lai, W., & Wade, L. (2020). Perceptual learning, talker specificity, and sound change. *Papers in Historical Phonology*, 5, 90–122. <https://doi.org/10.2218/pihph.5.2020.4439>
- Thorin, J., Sadakata, M., Desain, P., & McQueen, J. M. (2018). Perception and production in interaction during non-native speech category learning. *The Journal of the Acoustical Society of America*, 144(1), 92–103. <https://doi.org/10.1121/1.5044415>
- Tyler, M. D., & Best, C. T. (2024). The perceptual assimilation model: Early bilingual adults and developmental foundations. In M. Amengual (Ed.), *The Cambridge handbook of bilingual phonetics and phonology* (pp. 147–172). Cambridge University Press.
- Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, 167(3917), 392–393. <https://doi.org/10.1126/science.167.3917.392>
- Wrembel, M., Marecka, M., & Kopečková, R. (2019). Extending perceptual assimilation model to L3 phonological acquisition. *International Journal of Multilingualism*, 16(4), 513–533. <https://doi.org/10.1080/14790718.2019.1583233>
- Wright, R. (2001). Perceptual cues in contrast maintenance. In *The role of speech perception in phonology* (pp. 251–277). Brill.
- Wright, R. (2004). A review of perceptual cues and cue robustness. *Phonetically based phonology*, 34(57). Cambridge University Press.

